WP. 27
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (iv): Micro editing – methods and software

Redesigning a Data Editing Concept

Invited paper

Prepared by Elmar Wein, Federal Statistical Office, Germany[1]

I. Introduction

1. The German structural business statistics in domestic trade, accommodation and food service activities are based on two surveys. The data editing process is supported by an old mainframe application which has to be replaced by a new one. As there are many changes of the technical and organisational preconditions a complete redesign of the existing data editing concept was necessary.

2. The contribution describes first the objectives to be achieved by the new software and the basic conditions for redesigning the data editing concept. The second part of this document contains a description of the new data editing concept and selected features of the new software. The aim of that part is to contribute to best practices as regards (a software for) micro editing.

II. Considerations before the development of a new data editing concept

A. Objectives of the new data editing concept

3. The new IT-application and data editing concept should help to achieve the following aims:
   a) The effort for data editing and maintaining the data editing concept has to be reduced.
   b) The data quality of the structural business statistics should be maintained or be improved.

4. The two aims led to an intensive assessment of the …[i]
   - structural business statistics on one hand and the respondents on the other,
   - data that are collected and disseminated,
   - analysis of the existing data editing process including strengths and weaknesses of the old concept,
   - survey processes which adjoin the data editing process, and
   - resources like personnel, methodology, and information technology (IT).

5. The following sections will treat the aspects mentioned above. Statements in cursive letters will represent the conclusions regarding the redesign of the data editing concept.

B. Basic conditions for redesigning the existing data editing concept

6. The structural business statistics are projected sums, averages, and relations which are subdivided by the German states or detailed categories of the national version of the European classification of economic activities (NACE Rev. 2). Some results are classified on the level 3 of the NACE and up to five turnover classes. Besides these types of results anonymized micro data are disseminated, preliminary

[1] Phone: + 49 (0)611 75 3128

results are not disseminated.

7. The number of enterprises, affiliates, employees, the turnover, salaries, and gross profit are the most important variables. Other variables like the stock of trading goods and raw materials are only demanded from single users like the System of the National Accounts.[ii]
The dissemination of deeply categorised data and micro data induce higher demands on the data editing because there is a need to check all reasonable relations between the survey characteristics.

8. In addition to the statistics metadata on item and unit nonresponse and the statistics' precision are disseminated.
The new data editing process has to deliver metadata which will facilitate the computation of the quality reports. To fulfil the demand for disseminating metadata on item nonresponse a "slim" solution has to be found which do not increase the complexity of the new IT-system.

9. The analysis of the statistics and the data sources represent the most important basics for redesigning a data concept. German enterprises are permitted to transmit their annual balance sheets up to two years after the respective business year. As there is a difference of nearly two years between the reporting year of a structural business survey and the calendar year it is known that the enterprises complete the questionnaires on the basis of their balance sheets. An analysis of the duty for accounting revealed that there are significant simplifications for smaller enterprises with a maximum annual turnover of 500 000 EUR and a maximum annual profit of 50 000 EUR. These enterprises report to the tax authorities only the gross income and selected expenditures which are partly identical with the survey characteristics. Information on stock changes e.g. is completely missing.
The detailed analysis of the official tax forms revealed that nearly every enterprise is able to deliver data on affiliates, employees, turnover, payments for trading goods and salaries. These characteristics should form the set of core variables that determine the plausibility of other variables.

10. Hints on the plausibility of the data in the past were obtained by an analysis of the raw and plausible data. The following table shows the number of manual / automated corrections per survey characteristic of the raw data and the mean absolute correction:[iii]

No.	Survey characteristics	Absolute corrections	
		Number	Mean [%]
1.	Expenditures for services	1,322	103
2.	Paid taxes	557	594
3.	Trading goods, beginning of the year	508	240
4.	Turnover by wholesale trade (NACE)	494	69
5.	Number of employees	425	89
6.	Expenditures for trading goods	331	151
7.	Trading goods, end of the year	325	345
8.	Expenditures for commodities	321	6,501
9.	Part time employees	310	99
10.	Turnover by wholesale trade (NACE)	244	90

11. Variables that will not belong to the set of the core variables possess most of the corrections in wholesale trade statistics which are based on around 13 000 records. Two exceptions from this point of view are the number of the (part time) employees on the ranks 5 and 9. The variables on rank 4 and 10 are related to the coding of the enterprises' economic activities. The means of the ranks 1, 9, and 10 indicate the imputation of missing values because they are very close to 100 – the value that was set per definition when there was no raw date available.

12. As the previous table does not inform on the distributions of the survey characteristics an analysis of the percentiles combined with a graphical analysis (histograms combined with adequate theoretical distribution) and goodness-of-fit-tests were undertaken and used for developing imputation methods or acceptance limits of signals. The analysis of survey characteristics which are not well checked by the current data editing concept revealed the existence of outliers in published data.

The new data editing concept has to handle item nonresponse by offering modules for imputing. New checks / signals or automated corrections have to be introduced to detect or to replace outliers.

13. Big enterprises are obliged to create and publish their balance sheets via a German Internet platform. It is known that nowadays some of the respondents refuse to submit the questionnaire and refer to published data. It is expected that big enterprises will have the opportunity to transmit their balance sheets including their profit and loss accounts in a standardised format to an official authority in the medium term. It will be then the duty of a statistical office to use these data for the structural business statistics in spite of different meanings of some variables. These differences may induce a need for additional estimation methods.
A remarkable amount of electronic balance sheets will occur only in the medium term.

14. The wholesale and retail trade in Germany are characterised by a concentration of the enterprises with the respective Gini coefficients of 0.87 / 0.81 and around 0.65 for accommodation and food service activities.[2] The existing sample design takes advantage of this concentration because the most important enterprises – as regards their turnover – are arranged in census strata while the smaller ones are stratified in representative strata. Due to this sample design small enterprises may influence single statistics too because they possess weighting factors up to 50.
As regards data editing there is a need to flag the most important enterprises and to check them very carefully. The deep categorised results on one hand and the small enterprises with big weighting factors on the other may lead to the fact that even small enterprises may become very important. Consequently a mechanism to determine the most important enterprises has to take into account the weighting factors.

15. The two surveys of the structural business statistics in domestic trade, accommodation, and food service activities are highly harmonised because there are identical variables and sections in both questionnaires.
The standardised questionnaires shall be used for harmonising the specifications of checks and imputation methods. Consequently there will be common data editing modules besides specialised ones. A module refers to a section of a questionnaire and contains all respective checks. This decision will reduce the effort for maintaining the specifications.

16. The structural business statistics are based on one sample which is annually updated by an automated sample rotation. As a consequence around 17% of the enterprises are new and need an intensive checking of their economic activities. In addition to this the data of the structural business statistics are the basis for decisions as regards the participation of the enterprises in the surveys of the short term statistics.
The "clarifying function" of the structural business statistics in the case of new enterprises may require an intensive checking. This will lead to enterprise related data editing procedures.

17. The data are collected by paper and Internet questionnaires that contain only single checks on serious errors. Subject matter statisticians report that especially data entry errors are typical for the data submitted via Internet questionnaires.
First it should be checked the opportunity to complement the Internet questionnaires by additional checks. If this proposal cannot be realised additional checks should be included in the new data editing methodology.

18. The information of the respondents is collected by paper and Internet questionnaires. Some of the statistical offices scan the paper questionnaires and extract by the way the data, other offices let the data be captured, and the third want to capture data and clean them in one process. As a consequence the new software has to support the data entry as well as the data import.

19. During the data editing process a need for exporting a part of the data may occur for debugging the new software or improving the checks. It has to be ensured that the exported data cannot be used for the computing public statistics.

[2] The Gini coefficient is 1 in the case of a monopoly.

20. After finishing the data editing process one version of the plausible data is needed for detailed analysis with statistical standard software. This data have to contain all variables and aggregates so that the personnel can analyse it without writing program code.

21. In addition to this the plausible data has to be transformed in a special format so that the existing programs for computing the standard tables can be used.
The new software has to support the types of data import and export mentioned above.

22. The verification of the data can be facilitated by external data. The official statistics of Germany possess tax data of the enterprises, survey data used for short term statistics, and the data of the previous structural business statistics. An analysis of the turnover characteristics of the tax data (especially the tails of the empirical distributions) revealed bigger differences from the structural business statistics. The differences could be reduced by computing the turnover per employee. In the case of the short term statistics the differences were significantly smaller and a little bit larger in the case of the previous structural business statistics.
The occasional big differences between the tax data and the structural business statistics affect the verification of the data of the new enterprises. Opposed to that the relatively small differences between the short term statistics and the structural business statistics offer the opportunity to verify the turnover data of the respective enterprises and the data of the short term statistics can be used for imputing.

23. The survey in domestic trade consists of around 100 characteristics which are verified by 85 checks / deterministic imputations and 33 signals. Nearly one half of the checks are needed for clarifying the economic activity of an enterprise. Compared to that the data editing concept in accommodation and food service activities consists only of 27 checks and 18 signals. The great majority of the checks verify relations between the variables of a record. Checks regarding relations to the previous reporting period or external data from administrative sources are rarely.

24. There are only deterministic imputations available which bear hardly in mind neither the specific conditions of the different economic activities nor the different situations of small and big enterprises. If there is item nonresponse and no imputation method available the subject matter statisticians impute with the value of the previous year. Opposed to that there is an imputation method for the most important variable turnover available that uses the turnover information from the short term statistics. Unit non response is compensated by a very simple imputation method which creates a record with a minimal set of “conservative” values. Consequently the method induces under coverage and distorts the distributions of the variables.
The new imputation methods shall bear in mind the size of an enterprise and the specific aspects of its economic activity. The practice of imputing with values of the previous year has to be replaced because there are significant changes from one year to another. Unit non response shall be compensated by a reweighting approach.

25. After the termination of the data editing process a copy of the data is used for tabulating. The subject matter statisticians compare selected tables manually with the respective ones of the previous year to detect suspect results. At the end of the data editing process the statistical offices of the German states transmit their data to the Federal Statistical Office where subject matter statisticians perform the same manual comparisons of the tables. In general between one and three statistical offices have to reopen the data editing process every year for eliminating suspect data.
The manual and extensive comparisons of the tables shall be supported by an IT-module that documents the most suspect results together with the involved records. This tool should be available for the statistical offices of the states as well as for the Federal Statistical Office. It should be part of the new data editing process so that there will be no additional effort for tabulating.

26. The maintenance of the checks is very extensive because of their huge number and the fact that identical checks are implemented in different IT-systems. The realisation of the checks is also elaborately because subject matter statisticians first specify the checks which are then realised by IT-specialists. Finally the subject matter statisticians have to check the realisation.
Subject matter statisticians should be able to specify, test and integrate checks by their own in the new production environment.

27. The personnel of the statistical offices are responsible for up to 8 different surveys. Due to the on-going savings it is nowadays a common practice to suppress the great majority of the signals. In addition to this it is expected that budget cuts and savings of personnel will go on in the future.
The on-going savings require the introduction of automated data editing procedures. Besides this basic consequence signals have to be checked as regards their need and whether they can be transformed into automated corrections or checks. If signals are still necessary their hit rates should be improved. In addition to this error messages should possess explanatory information like indicators together with the data which initiate a check / signal.

28. The old software for data editing was used about 13 years and it is expected that the new one will be used for the same duration. The new IT-infrastructure of the decentralised German statistical system supports a central database that will contain around 65 000 records per year. It permits access to all statistical offices via secured Internet connections. The statistical standard software SAS will be more and more available in the next few years among all statistical offices of the German states which are responsible for the data editing processes.

29. As regards more powerful data editing methods there is only a SAS macro available that supports a selective editing.[iv] SAS macros for an automated determination of erroneous variables are still missing, SAS macros for imputing are currently under development. Besides this there is a SAS prototype available that supports a comparison of statistical results with the ones of the previous year.
The storage of the whole sample in one central database sets good preconditions for automated data editing methods. The enormous effort for the development of the new IT-application on one hand and its expected long useful life on the other should lead to the integration of methods and tools like SAS. As there is no tool for determining erroneous variables available the introduction of automated data editing methods will be adjourned to one of the following years. This bottleneck determines mainly the data editing process: As all errors have to be corrected manually on one hand and statistical standard results and anonymized micro data are only required on the other it is at the moment not useful to introduce selective editing methods in a complete new data editing application. To prepare the subject matter statisticians for macro editing the computation of "TOP-enterprises" and an automated comparison of annual results should be introduced where the last method has to compensate the methodological weakness of micro editing.

30. Besides the software for supporting data editing there is a software available that enables subject matter statisticians to *program* survey characteristics and checks, test and to integrate them autonomously in the production environment.[v]
Subject matter statisticians who were responsible for specifying the checks in former times now have to be familiar with simple programming techniques and must have the ability to debug source code.

C. Achieving the aims

31.The main aim of the new data editing concept is to reduce the effort for data editing. Bearing in mind the preconditions mentioned above this aim can be achieved at the moment only by …[vi]
- introducing an automated comparison of actual results with the ones of the previous year,
- expanding the imputations and automated corrections on micro level based on a "hierarchy of characteristics",
- removing the number of worthless signals and introducing new checks so that inconsistencies are discovered in an early phase of the data editing process,
- improving the hit rate of plausibility checks by the use of enterprises' specific reference values.

32. It is expected that the four changes will also improve the data quality. The planned reweighting approach is the largest contributor to this aim, followed by improving the hit rate of the plausibility checks and the introduction of new checks.

33. The efforts for maintaining the data editing concepts will be reduced by …
- a consequent harmonisation of the checks and survey characteristics,
- the ability of the subject matter unit to specify, test and implement the imputations and checks autonomously.

III. Selected aspects of the new data editing concept and the software

A. Determination of TOP-enterprises

34. The determination of TOP-enterprises shall ensure that the subject matter statisticians are aware of the most important units. In addition to this the criterion “TOP-enterprise” can be used to decide on the employment of the subject matter statisticians. When advanced data editing methods will be available the criterion “TOP-enterprise” will be used for determining the units which have to be checked manually.

35. TOP-enterprises are defined as units with biggest weighted turnover which cover x per cent of the weighted turnover of a chosen NACE class and a German state. Referring to the dissemination policy of the statistical offices of the German states the NACE level 3 was chosen. The determination of the enterprises requires that the administrator of a statistical office fixes the percentage rate first in a key file. This approach is very flexible and thus supports the dissemination policy of a statistical office in a best way.

36. The documentation of TOP-enterprises consists of the following reports:
- the number of enterprises per NACE class and the individual shares of the weighted turnover,
- a summary of the data editing as regards the TOP-enterprises, and
- a list with the TOP-enterprises with missing data.

B. Coding

37. One main task of the subject matter statisticians is to clarify the economic activity - especially of the new enterprises. The task represents the beginning of the data editing process. The clarification will be performed by checks of the relations between the NACE code and the turnover in per cent categorised by groups of traded goods.

38. The checks refer to a key file which contains valid combinations of codes and percentages of turnover categorized by traded goods. If there is a change of the classification of economic activities all enterprises obtain codes of the old and new classification. Consequently the key file can be supplemented with the valid combinations of old and new codes. Further assistance for coding the economic activity of an enterprise will be provided by links to the …:
- central classification server of all statistical offices
  It offers the opportunity to retrieve the economic activity by a keyword based search and contains all valid codes of economic activities.
- German electronic trade register
  It provides information on enterprises like the number of affiliates and the economic activity.
- homepage of a German commercial database on enterprises
  An additional access to information on enterprises.
- homepage of the Google search
  To be used for searching an enterprise’s Internet site.

The connection to the central classification server will be improved within the next two years that means the server will provide Web Services which will facilitate the coding.

C. Data editing

39. The new software stores the plausible data as well as the respective raw data. The user will be able to call for both types of data of the current reporting period and the data of the last two years. In addition to this a record contains a text field for comments on (the data of) an enterprise and additional fields may facilitate the data editing of a record, e.g. information on commission trade in retail trade or the existence of a short business year.

40. The new data editing concept is valid for both surveys because it consists of common modules with checks and imputation methods and three common key files:
- Data related to economic activities (see section E).

- Key data for data editing
  They contain two types of data: type I represents limits and type II consists of basic data for imputations like relations, proportions or components of estimation models. The data are linked to checks and records of the enterprises via the unique identification numbers of the checks, the code of the economic activity (optional), the region (optional, east and west of Germany, administrative regions), and the turnover category.
- Data related to powerful editing methods
  The current version of this key file is considered as experimental. It contains only the minimal coverage of turnover per NACE class and German state.

The number of key files was limited to restrict their maintenance. The limitation was achieved by an integration of different types of data in one file structure.

41. The new software supports two types of data input: data entry and data import. It will offer two courses of data editing balanced to the type of user: one course aims at detecting all data entry errors because it checks only the ranges of the data and the second course checks all types of errors.
The software used for data editing supports the checking of one record as well as the checking of all existing records.

42. The course of data editing which aims at detecting all errors first checks the key characteristics like the code of the economic activity, the turnover, the personnel, and the code of the region (east / west of Germany). This is an important step because an enterprise is associated to homogenous peer groups on the basis of these characteristics. A preparing analysis showed numerous and significant correlations between the turnover and other survey key characteristics. An analysis of the variance revealed that the code of the economic activity on the NACE level 3 and 14 classes of turnover lead to homogenous peer groups with sufficient number of cases and minimal variances of important related variables. The assignment of an enterprise to a homogenous peer group is considered to be a prerequisite for plausible estimations which will pass the checks.

43. A classical micro editing of a record is suitable for detecting and correcting obvious implausible relations between different survey characteristics. One typical example is “a retailer with no store”. One weakness of the approach is that is hard to detect values that seem to be unlikely. A typical example of this problem is “the number of stores is different at the beginning of the year in 20 per cent from the number of stores from the previous year”. This is a weak approach because this signal flags a record with 1 store in the past and 2 stores in the present which may be plausible. As the practice shows that such signals were more and more ignored it was decided to establish relative limits in per cent depending on the size of the respective enterprise in the previous business year as shown in the following table (example):[vii]

A	B	C	D
stores in the previous year	digits	relative limit in per cent (100-10*digit(A)-A/10^(digit(A)-1)	rounded(absolute tolerance limit)
1	1	89	1
9	1	81	7
10	2	79	8
99	2	70	69
100	3	69	69
999	3	60	599
1000	4	59	590
9999	4	50	5000
10000	5	49	4900

Referring to the example first the need was checked to establish limits for the peer groups or not. After clarifying this question it was decided to restrict the effort for maintaining the planned signal by establishing a function which maps the basic date (here the number of stores in the previous year) on the tolerance limit in per cent. The critical values in per cent were obtained by an analysis of the deviations from the last two surveys. On the basis of the observed empirical distribution the needed function (C) was developed which meets the most important percentiles approximately (D).

44. The additional efforts mentioned in the previous section are only able to restrict an under or over coverage of a survey characteristic to a limited extent. To improve the detection of these effects the

record of an enterprise contains besides the survey characteristics important relations of them like the gross profit per local unit, the cost of sales or the annual gross salary per employee.

45. The micro editing will contain imputation methods. Deterministic imputation methods were implemented as well as methods based on relations observed (for the peer groups). An imputation method can be executed automatically in the case of item non response or after a command of a subject matter statistician. The second alternative represents an automated correction. It was realised for editing a complex section of the questionnaire that consists of a matrix with 39 rows and 3. It is assumed that a subject matter statistician will decide on the basis of a visual inspection of the matrix and the differences between the entries in the matrix and the totals of an enterprise. To support the decision a check delivers the differences between both data. To restrict the effort for the maintenance both methods are based on the same module but the automated correction does not possess an error detection module. In general both methods list at the end the changed characteristics and the imputed / corrected data are always verified by a corresponding check.

46. While examining suspicious statistical results the selection of the causative records plays an important role. The user of the new software will be able to select records by the identification number of an enterprise, the identification number of activated checks, the plausibility status of the records, the status "old/new enterprise", the NACE code as well as the code of a check, and the status "TOP-enterprise". The criteria can be combined by the logical "And-operator".

47. The reporting of data editing is available for users who are responsible for the data editing of an assigned number of enterprises as well as for users who coordinate the work of the subject matter statisticians. To restrict the number of reports the following aims were defined first:

a) Documenting the partition of the enterprises among several subject matter statisticians.
   With the help of this basic report it shall be verified that all enterprises of the two surveys are distributed equally among the personnel.
b) Providing an overview of the editing (of TOP-enterprises) at a point in time.
   At the end of a data editing process the individual burden of the subject matter statisticians may be different due to different numbers of errors and different developments of the incoming data. The respective report is available for a statistician as well as for the coordinator of a data editing process.
c) Observing the degree of unit non response.
   This report is considered to be important because new enterprises first have to take part at the structural business statistics. As a consequence the report informs on the willingness of new enterprises to cooperate in the medium term.
d) Documenting the number of corrections per survey characteristic and the difference between the respective weighted raw and plausible data.
   Raw and plausible data are the basis of this report. The corrections and imputations are documented by the differences between the raw and the plausible data in per cent of the plausible data. The report shall deliver hints for an optimisation of the data editing process.
e) Documenting the number and type of activated checks.
   This type of report shall supplement the previous report.

D. An automated biennial comparison of statistical results

48. Micro editing is suitable for detecting errors but possess uncertainties as regards observations that seem to be unlikely. Often suspicious observations are flagged by signals but nevertheless accepted on record level and consequently the respective aggregate suffers from under or over coverage. To compensate the deficit of micro editing a tool was developed which compares weighted plausible aggregates of the previous reporting period with the respective (implausible) ones of the actual reporting period on macro level. The aim of the tool is to detect under or over coverage and to facilitate the detection of the causative suspicious records.[viii]

49. As regards the choice of the aggregates several alternatives seem to be reasonable: Comparisons of homogenous peer groups would facilitate the detection of suspect records. A disadvantage of this processing is that the under or over coverage within the peer groups may differ from the effects of the

results which have to be disseminated.

50. Opposed to that comparisons of the results represent a more output oriented approach. In this case there are two aspects relevant: the necessity to ensure the detection of the suspicious records on one hand and the requirement to compute the results on sufficient numbers of records so that the figures to be compared are stable enough.

51. As the structural business statistics are disseminated by the German states with their small samples on the NACE level 3 this type of aggregate was chosen for the biennial comparisons. The categorisation of the results solely on the economic activity complicates the detection of the suspicious records.

52. In general the comparison of aggregates from different reporting periods may be biased which may lead to an erroneous priority setting as regards necessary corrections. The bias may be caused by …:

a) structural influences represented by different numbers of stores or employees
   An ANOVA of the structural business statistics reveals a stronger relation between employees and turnover than between the number of the stores and turnover. Unfortunately more and more German enterprises assign their personnel to enterprises with borrowed workforces which are not captured by a legal survey characteristic. Consequently the weighted number of stores is used for eliminating the structural influence.
b) the influence of the business cycle
   They may lead to higher / lower revenues or expenditures per enterprise. The elimination of these effects is simple in the case of plausible data by subtracting the mean from the individual observations. As the means of the suspicious data of the actual reporting period may be biased robust statistics of location will be used, e.g. the trimmed / winsorized mean or the median.
c) the natural variability of the survey characteristics
   Survey characteristics with bigger variances will differ more between two reporting periods and thus indicate alleged differences. To avoid random effects the plausible dataset may be cleaned by the variance and the dataset of the current reporting period by robust measures of variability like the mean absolute deviation, the interquartile range or a “trimmed variance”.

53. A biennial comparison is meaningful for important survey characteristics which are regularly observed. As the structural business statistics contain numerous variables combinations of them will be used like the number of local units per enterprise, the costs (salary + social insurance contribution) per employee, the gross profit ratio, and the value added at factor cost. An advantage of this approach is that relations between variables are implicitly checked.

54. Differences observed for the results of the characteristics are computed in per cent of the plausible results of the previous year so that one indicator can be computed per statistical result. Typical indicators for differences of a result are the sum, the mean, median or maximum of the differences per weighted variable in per cent.

55. The characteristics which are used on macro level are used on micro level too and they are standardised so that outliers at the tails of the empirical distributions can be easily detected. To indicate the contribution of a record to the aggregate the variables are weighted.

56. The methodology described above was realized by a collection of macros on the basis of SAS 9.2 with the additional packages SAS-Stat and SAS-Graph.[3] After the data import statistics’ specific operations can be performed via the inclusion of specific SAS program code. After these computations the macros support the different types of standardization mentioned in section 32c. At the end the macros list the results with the biggest differences on top and the respective records of the enterprises. An overview of the differences is provided by a histogram. The macros are commented in English, the text of the histogram can be customized via a language file.

[3] The SAS-macros were developed by Pascal Avieny.

IV. Conclusions

57. A comprehensive analysis of the preconditions has to be the first step for redesigning a data editing concept. Important aspects are the data and metadata to be delivered, the respondents with their data sources, (external) reference data, interfaces to existing survey processes, knowledge about the strengths and weaknesses of the existing data editing concept, empirical distributions of the survey characteristics, available IT and methodology and the existing personnel. Besides the data and metadata the existing personnel, the available IT, and the methodology seem to be the most important factors.

58. The availability of methods for an automated error detection and imputation is a decisive precondition as regards the introduction of a highly automated data editing process. Selective editing without the availability of these methods is not meaningful if there is no need to disseminate preliminary results but plausible anonymized micro data.

59. If the automated and selective editing methods are not available the redesign of a data editing concept can lead to improvements as regards the detection of the most influential units, the computer assisted micro editing and the introduction of a macro view via an automated comparison of actual aggregates with the respective plausible ones of the previous reporting period.

60. The determination of TOP-enterprises (=most influential units) in the absence of selective and automated editing methods is meaningful because it helps subject matter statisticians to recognize the important units. The most influential units are the ones with the biggest contributions to a publication cell. Their determination should be based on the most important variable which is the (weighted) turnover in the case of structural business statistics.

61. The effort for data editing in the absence of powerful methods can be limited by …

a) reducing the manual clarifications on the most important survey characteristics which have to be derived from a preparatory analysis,
b) the provision of automated imputation / correction methods for complex parts of a questionnaire,
c) a careful use and design of signals, and
d) adequate reports which support the management and optimisation of a data editing process.

62. Micro editing reaches methodological limits in the case of signals. This type of verification is necessary as long as there are survey characteristics with uncertain empirical distributions. Comparisons of plausible aggregates from the previous period with the one of the current period may help to detect under and over coverage on one hand and may restrict the effort for detecting the causative records when the differences on the macro level are combined with the respective records.

---

[i] Mark Pierzchala: “A review of the state of the art in automated data editing and imputation”, Statistical Data Editing, Vol.1: Methods and techniques 1, pp. 10-17, www.unece.org/stats/publications/editing/SDE1chA.pdf
Elmar Wein: “The planning of data editing”, pp. 4-5, Work session on statistical data editing, 18-20 October 2000, Cardiff, www.unece.org/stats/documents/2000/10/sde/3.e.pdf

[ii] Orietta Luzi, Ton De Waal, Beat Hulliger et al: “Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys”, pp. 10,
epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf

[iii] Elmar Wein: “The usability of corrections for improving and pricing data quality”, pp. 2-5, Work session on statistical data editing, 21–23 April 2008, Vienna, www.unece.org/stats/documents/2008/04/sde/wp.16.e.pdf

[iv] Elmar Wein: “Concepts, materials, and IT modules for data editing of German statistics”, pp. 7, Work session on statistical data editing, 16-18 May 2005, Ottawa, www.unece.org/stats/documents/2005/05/sde/wp.37.e.pdf

[v] Elmar Wein: “Concepts, materials, and IT modules for data editing of German statistics”, pp. 5, Work session on statistical data editing, 16-18 May 2005, Ottawa, www.unece.org/stats/documents/2005/05/sde/wp.37.e.pdf

[vi] Elmar Wein: “Introducing and Implementing a new data editing strategy“, pp. 12-13, Work session on statistical data editing, 16-18 May 2005, Ottawa, www.unece.org/stats/documents/2005/05/sde/wp.14.e.pdf

[vii] UN ECE: “Glossary of terms on statistical data editing“, definition of “statistical edit”, p. 11, www.unece.org/stats/publications/editingglossary.pdf

[viii] Leopold Granquist: “Macro-Editing – The Aggregate method”, Statistical Data Editing, Vol.1: Methods and techniques 1, pp. 137, http://www.unece.org/stats/publications/editing/SDE1.htm